STA 360/602L: Module 4.5

Missing data and imputation I

Dr. Olanrewaju Michael Akande

Introduction to missing data

- Missing data/nonresponse is fairly common in real data. For example,
  - Failure to respond to survey question
  - Subject misses some clinic visits out of all possible
  - Only subset of subjects asked certain questions
- Recall that our posterior computation usually depends on the data through $p(Y|\theta)$, which cannot be computed (at least directly) when some of the y_i values are missing.
- The most common software packages often throw away all subjects with incomplete data (can lead to bias and precision loss).
- Some individuals impute missing values with a mean or some other fixed value (ignores uncertainty).
- As you will see, imputing missing data is actually quite natural in the Bayesian context.

Missing data mechanisms

- Data are said to be missing completely at random (MCAR) if the reason for missingness does not depend on the values of the observed data or missing data.
- For example, suppose
  - you handed out a double-sided survey questionnaire of 20 questions to a sample of participants;
  - questions 1-15 were on the first page but questions 16-20 were at the back; and
  - some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be MCAR if they simply did not realize the pages were double-sided; they had no reason to ignore those questions.
- **This is rarely plausible in practice!**

Missing data mechanisms

- Data are said to be missing at random (MAR) if, conditional on the values of the observed data, the reason for missingness does not depend on the missing data.
- Using our previous example, suppose
  - questions 1-15 include demographic information such as age and education;
  - questions 16-20 include income related questions; and
  - once again, some participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be MAR if younger people are more likely not to respond to those income related questions than old people, where age is observed for all participants.
- **This is the most commonly assumed mechanism in practice!**

Missing data mechanisms

- Data are said to be missing not at random (MNAR or NMAR) if the reason for missingness depends on the actual values of the missing (unobserved) data.
- Continuing with our previous example, suppose again that
  - questions 1-15 include demographic information such as age and education;
  - questions 16-20 include income related questions; and
  - once again, some of the participants did not respond to questions 16-20.
- Then, the values for questions 16-20 for those people who did not respond would be MNAR if people who earn more money are less likely to respond to those income related questions than old people.
- **This is usually the case in real data, but analysis can be complex!**

Mathematical formulation

- Consider the multivariate data scenario with $\boldsymbol{Y}_i = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^T$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T$, for $i = 1, \ldots, n$.
- For now, we will assume the multivariate normal model as the sampling model, so that each $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$.
- Suppose now that $\boldsymbol{Y}$ contains missing values.
- We can separate $\boldsymbol{Y}$ into the observed and missing parts, that is, $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$.
- Then for each individual, $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis})$.

Mathematical Formulation

- Let
  - j index variables (where i already indexes individuals),
  - $r_{ij} = 1$ when y_{ij} is missing,
  - $r_{ij} = 0$ when y_{ij} is observed.
- Here, r_{ij} is known as the missingness indicator of variable j for person i.
- Also, let
  - $\boldsymbol{R}_i = (r_{i1}, \ldots, r_{ip})^T$ be the vector of missing indicators for person i.
  - $\boldsymbol{R} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)$ be the matrix of missing indicators for everyone.
  - $\boldsymbol{\psi}$ be the set of parameters associated with $\boldsymbol{R}$.
- Assume $\boldsymbol{\psi}$ and $(\boldsymbol{\theta}, \Sigma)$ are distinct.

Mathematical Formulation

- MCAR:

$$p(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\boldsymbol{R}|\boldsymbol{\psi})$$

- MAR:

$$p(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi})$$

- MNAR:

$$p(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = p(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi})$$

Implications for likelihood function

- Each type of mechanism has a different implication on the likelihood of the observed data $\boldsymbol{Y}_{obs}$, and the missing data indicator $\boldsymbol{R}$.
- Without missingness in $\boldsymbol{Y}$, the likelihood of the observed data is

$$p(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \Sigma)$$

- With missingness in $\boldsymbol{Y}$, the likelihood of the observed data is instead

$$p(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = \int p(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \Sigma)\mathrm{d}\boldsymbol{Y}_{mis}$$

- Since we do not actually observe $\boldsymbol{Y}_{mis}$, we would like to be able to integrate it out so we don't have to deal with it.
- That is, we would like to infer $(\boldsymbol{\theta}, \Sigma)$ (and sometimes, $\boldsymbol{\psi}$) using only the observed data.

Likelihood function: MCAR

- For MCAR, we have:

$$
\begin{aligned}
p(\boldsymbol{Y}_{obs}, \boldsymbol{R} | \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) &= \int p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\
&= \int p(\boldsymbol{R} | \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\
&= p(\boldsymbol{R} | \boldsymbol{\psi}) \cdot \int p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\
&= p(\boldsymbol{R} | \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs} | \boldsymbol{\theta}, \Sigma).
\end{aligned}
$$

- For inference on $(\boldsymbol{\theta}, \Sigma)$, we can simply focus on $p(\boldsymbol{Y}_{obs} | \boldsymbol{\theta}, \Sigma)$ in the likelihood function, since $(\boldsymbol{R} | \psi)$ does not include any $\boldsymbol{Y}$.

Likelihood function: MAR

- For MAR, we have:

$$\begin{aligned} p(\boldsymbol{Y}_{obs}, \boldsymbol{R} | \boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) &= \int p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\ &= \int p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\ &= p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \cdot \int p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis} \\ &= p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs} | \boldsymbol{\theta}, \Sigma). \end{aligned}$$

- For inference on $(\boldsymbol{\theta}, \Sigma)$, we can once again focus on $p(\boldsymbol{Y}_{obs} | \boldsymbol{\theta}, \Sigma)$ in the likelihood function, although there can be some bias if we do not account for $p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{\psi})$, since it contains observed data.

- Also, if we want to infer the missingness mechanism through $\boldsymbol{\psi}$, we would need to deal with $p(\boldsymbol{R} | \boldsymbol{Y}_{obs}, \boldsymbol{\psi})$ anyway.

Likelihood function: MNAR

- For MNAR, we have:

$$p(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \Sigma, \boldsymbol{\psi}) = \int p(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) \cdot p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \Sigma) \mathrm{d}\boldsymbol{Y}_{mis}$$

- The likelihood under MNAR cannot simplify any further.
- In this case, we cannot ignore the missing data when making inferences about $(\boldsymbol{\theta}, \Sigma)$.
- We must include the model for $\boldsymbol{R}$ and also infer the missing data $\boldsymbol{Y}_{mis}$.

How to tell in practice?

- So how can we tell the type of mechanism we are dealing with?
- In general, we don't know!!!
- Rare that data are MCAR (unless planned beforehand); more likely that data are MNAR.
- **Compromise**: assume data are MAR if we include enough variables in model for the missing data indicator $\boldsymbol{R}$.
- Whenever we talk about missing data in this course, we will do so in the context of MCAR and MAR.

What's next?

Move on to the readings for the next module!